

Homophilic Clustering by Locally Asymmetric Geometry

Deli Zhao

DELI_ZHAO@HTC.COM

HTC Beijing Advanced Technology and Research Center, Beijing, China

Xiaoou Tang

XTANG@IE.CUHK.EDU.HK

Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong

Abstract

Clustering is indispensable for data analysis in many scientific disciplines. Detecting clusters from heavy noise remains challenging, particularly for high-dimensional sparse data. Based on graph-theoretic framework, the present paper proposes a novel algorithm to address this issue. The locally asymmetric geometries of neighborhoods between data points result in a directed similarity graph to model the structural connectivity of data points. Performing similarity propagation on this directed graph simply by its adjacency matrix powers leads to an interesting discovery, in the sense that if the in-degrees are ordered by the corresponding sorted out-degrees, they will be self-organized to be homophilic layers according to the different distributions of cluster densities, which is dubbed the Homophilic In-degree figure (the HI figure). With the HI figure, we can easily single out all cores of clusters, identify the boundary between cluster and noise, and visualize the intrinsic structures of clusters. Based on the in-degree homophily, we also develop a simple efficient algorithm of linear space complexity to cluster noisy data. Extensive experiments on toy and real-world scientific data validate the effectiveness of our algorithms.

1. Introduction

Clustering is a fundamental task in machine learning, computer vision, information retrieval, and data mining. Data generated in practice generally possess the property of lo-

The original work was performed when the first author worked in Department of Information Engineering, the Chinese University of Hong Kong in 2012. The paper was revised and re-submitted when he worked in HTC. The paper was rejected by Nature in 2012, NIPS in 2013, and ICML in 2014. The permanent email of the first author: zhaodeli@gmail.com.

cal or global aggregation in sample spaces due to pattern correlations; this lays the foundation of detecting clusters in data points.

The classic algorithms of clustering are k-means and the hierarchical agglomerative algorithms based on linkages, such as the single, average, and complete linkages. The k-means algorithm iteratively optimizes clusters by minimizing distance squares between the center of each cluster and associated cluster members, which is simple and easily usable. The distance-based partitional method is put forward by the algorithm of Affinity Propagation (AP) (Frey & Dueck, 2007) that is proven to be fast only with simple manipulations of sparse networks. The most popular linkage algorithms is the average linkage, which measures the structural proximity of pairwise clusters by the arithmetic mean of distances between all members in the two clusters. The framework of hierarchical agglomerative clustering is also applied in the advanced graph-theoretic models, such as graph cycles (Zhao & Tang, 2008) and directed linkages (Zhang et al., 2012).

In addition to the above conventional frameworks, spectral clustering is a different type of approaches that can cluster data of complex structures. For example, the Normalized Cuts (NCuts) method hierarchically splits data by the graph Laplacian in the divisive way (Shi & Malik, 2000). Alternatively, the k-means or other types of clustering algorithms can be performed on spectral coordinates derived by eigenvectors of graph-Laplacian matrix (Ng et al., 2001; Meilă & Pentney, 2007; Zhou et al., 2005). Spectral embeddings of Laplacian can unfold the underlying manifolds in low-dimensional spaces (Belkin & Niyogi, 2003). Therefore, spectral clustering is free from the limitation of data structures or distributions. A variant of requiring low-dimensional coordinates was presented by (Lin & Cohen, 2010), which is based on matrix power iterations. However, these algorithms will encounter difficulty when clustering data contaminated with noise. Another type of clustering algorithms that have been proven to be noise-robust are based on the application of graph kernels or their ana-

logues. The entries of graph kernel matrices can be viewed as the measurement of global similarities between data points. With the kernel-enhanced similarities, the proximal correlations between data points can be more accurately measured. The hierarchical clustering algorithms are usually employed on graph kernels, such as the matrix power kernel (Newman & Girvan, 2003; Dongen, 2000), the von Neumann kernel (Katz, 1953), and the diffusion kernel (Kandola et al., 2003). The obvious limitation of such algorithms is that the space complexity is square in the number of data points.

We are now in the era of data deluge. The large scales of data incur two difficulties for data clustering. Firstly, the space complexity of algorithms should be sufficiently low that the available RAM is adequate to run the algorithms. Secondly, the large-scale data usually contains noise or outliers, thereby requiring the algorithms to identify outliers and noise. These issues necessitate the development of clustering algorithms that are robust to noise or outliers with low space complexity.

In this paper, we propose new algorithms to address the issue of accurately clustering noisy data with low complexities of space and time. Our algorithmic framework is based on an intriguing property of directed graphs drawn from data. The asymmetries of the local neighborhoods of each data point lead to a directed graph that is embedded in high-dimensional space. We discover that the arrangement of high-order in-degrees ranked by corresponding sorted out-degrees on such directed graphs breeds the homophilic distribution of data points according to different densities. This density homophily classifies data points into transparent layers according to the values of in-degrees. Noisy data or outliers have low densities such that they are aggregated to form the weakest layer, making it easy to find the boundary between clusters and noise. In addition, the cores of all clusters can be singled out simultaneously by the ratios of in-degrees to out-degrees, thereby greatly facilitating the performance of clustering noisy data. Based on density homophily, we develop a simple algorithm for clustering. Our algorithm passes similarities with local connectivity of directed graphs according to homophilic priority. It attains the better accuracy of clustering with the linear space complexity while maintaining the low time complexity.

2. Density Homophily with Digraph

Homophily is a concept that describes the behavioral preference of individuals with others who have similar attributes to themselves in social sciences. These attributes include age, gender, race, belief, interests, etc. The conventional work in the seminal paper (McPherson et al., 2001) and a recent one (Kossinets & Watts, 2009) comprehensively studied the homophilic organization in so-

cial networks. It has been recently reported that besides popularity modeled by power laws of degree distributions (Barabási & Albert, 1999), homophily is another dimension of characterizing the preferential attachment of new links in real evolving networks (Papadopoulos & Kitsak, 2012). We find that high-order in-degrees in geometric digraph show the transparent homophilic distributions with similarity propagation. These homophilic distributions are associated with different densities of clusters in data points.

2.1. Neighborhood Asymmetry and Digraph Model

Suppose that a data set $\mathcal{X} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{M}^d, i = 1, \dots, n\}$ is provided, where \mathcal{M}^d is the d -dimensional sample space and n is the number of data points. \mathcal{M}^d may be the mixture of manifolds or multivariate Gaussians. For an arbitrary data point \mathbf{x}_i , we may search its k nearest neighbors (NNs) $\mathcal{N}_i = \{\mathbf{x}_{i_p} | \mathbf{x}_{i_p} \in \mathcal{M}^d, p = 1, \dots, k\}$ with respect to a pre-defined distance metric. Assume that another data point \mathbf{x}_j is one of NNs of \mathbf{x}_i , say, $\mathbf{x}_j \in \mathcal{N}_i$. For the set \mathcal{N}_j of NNs of \mathbf{x}_j , there are two possible cases for the structural relationship of \mathbf{x}_i to \mathbf{x}_j : $\mathbf{x}_i \in \mathcal{N}_j$ or $\mathbf{x}_i \notin \mathcal{N}_j$. In other words, \mathbf{x}_i may not necessarily be the NN of \mathbf{x}_j if \mathbf{x}_j is the NN of \mathbf{x}_i . This neighborhood asymmetry is the most fundamental fact of spatial adjacency on local neighborhoods of data cloud. If we locally connect data points by a graph \mathcal{G} of NNs, a weighted adjacency matrix can be formed by

$$W_{i \rightarrow j} = \begin{cases} \text{sim}(\mathbf{x}_i, \mathbf{x}_j), & \text{if } \mathbf{x}_j \in \mathcal{N}_i \\ 1, & \text{if } j = i \\ 0, & \text{otherwise} \end{cases}, \quad (1)$$

where i successively goes from 1 to n , meaning that the graph \mathcal{G} is constructed row by row. $\text{sim}(\mathbf{x}_i, \mathbf{x}_j)$ presents the value of similarity between \mathbf{x}_i and \mathbf{x}_j , and $\text{sim}(\mathbf{x}_i, \mathbf{x}_j) \in [0, 1]$. The similarity measurements may be the cosine value, the distance exponential such as $e^{(-d_{ij}^2/\sigma^2)}$, where σ is a free parameter, or other variants of similarity measures. $W_{i \rightarrow j}$ is the (i, j) -th entry of the weighted adjacent matrix \mathbf{W} of graph \mathcal{G} . Here we use $i \rightarrow j$ to emphasize that the link from \mathbf{x}_i to \mathbf{x}_j is directed, thereby forming a digraph \mathcal{G} .

Thus for an arbitrary node i , there are two types of structural measures: out-degree and in-degree. The out-degree is the sum of weights of out-going links from node i to its neighbors and the in-degree is the sum of weights of incoming links from neighbors pointing to node i . In matrix form, the out-degree vector \mathbf{d}^{out} of all nodes can be written as $\mathbf{d}^{\text{out}} = \mathbf{W}\mathbf{1}$, where $\mathbf{1}$ is the all-one vector of length n , and $\mathbf{d}^{\text{in}} = \mathbf{W}^\top \mathbf{1}$, where \top denotes the matrix transpose. It suffices to note that each node in \mathcal{G} is imposed with a loop of weight 1, thereby excluding the case of the vanishing in-degrees and out-degrees. The out-degrees and in-degrees

are the most elementary ingredients in the characteristics of complex networks.

2.2. Similarity Propagation

From the viewpoint of paths in \mathcal{G} , the structural connectivity modeled by \mathbf{W} can be regarded as the linkage of paths of length 1. Many studies have verified that long paths are favorable for modeling complex structures. For instance, the shortest paths are applied to characterize manifold and network structures (Tenenbaum et al., 2000). Long cycles can convey the high-level information of balance in signed networks (Zhao & Tang, 2008; Chiang et al., 2011). With long paths, the membership affinities within an arbitrary network community can be more accurately enhanced (Katz, 1953; Newman & Girvan, 2003). Similarity propagation by walks is the simplest and most intuitive of the various applications of paths. It can be written simply by matrix power \mathbf{W}^t , where t presents the length of paths to be investigated. The (i, j) -th entry $W_{i \rightarrow j}^t$ of \mathbf{W}^t can be interpreted as a kind of accumulative similarity between x_i and x_j by passing similarities in digraph \mathcal{G} by t steps. To make it clear, we expand $W_{i \rightarrow j}^t$ by graph representation to give

$$W_{i \rightarrow j}^t = \sum_{\{\text{all possible paths of length } t\}} \prod_{\{\text{one of paths of length } t\}} W_{i \rightarrow j}. \quad (2)$$

From (2), it is easy to see that $W_{i \rightarrow j}^t$ is essentially a global sum-product similarity generated by all possible paths of length t that connect node i and node j . The path-based similarity can capture the structural correlation of deep connections between data points. If we regard each data point as a human individual and the whole data set as the society, the growth of \mathbf{W}^t can be viewed as the dynamic process of individual social interactions. Therefore, we can apply the social concepts and principles for data analysis. The homophily pertaining to data we underscore is such a social property of data points.

With \mathbf{W}^t , we can define the t -order degrees as $\mathbf{d}^{in} = \mathbf{W}^t \mathbf{1}$ and $\mathbf{d}^{out} = (\mathbf{W}^t)^\top \mathbf{1}$. For convenience of representation, we have omitted the scripts of ‘ t ’ in \mathbf{d}^{in} and \mathbf{d}^{out} . It may be inferred from context. It is computationally prohibitive to directly compute \mathbf{W}^t for a large n , because \mathbf{W}^t turns to be a fully dense matrix for a moderate t . Actually, \mathbf{d}^{in} and \mathbf{d}^{out} can be derived by sparse-matrix-vector products iteratively. To maintain the values of \mathbf{d}^{in} and \mathbf{d}^{out} , we perform the sum-to-one normalization during iteration. The procedures are provided in Algorithm 1. Notice that we use the same scale constant to normalize \mathbf{d}^{in} and \mathbf{d}^{out} in each iteration in Algorithm 1. Such manipulation is crucial for the usage of in-degrees and out-degrees, which will be presented in the following section.

2.3. In-degree Homophily

An interesting property of the t -order \mathbf{d}^{in} and \mathbf{d}^{out} is that the in-degrees are self-organized to be homophilic layers if ordered by the associated out-degrees. The in-degrees reflect the popularity of nodes in digraph \mathcal{G} (Barabási & Albert, 1999; Papadopoulos & Kitsak, 2012), thereby differentiating the cluster densities of data points. In this way, the density distribution of clusters may be accurately visualized, providing a powerful avenue for intuitively analyzing clusters. We present the specific steps of illustrating the homophilic in-degrees (HI) in Algorithm 2. For simplicity, we call the visualization the HI figure.

Examples of the HI figure on toy data are shown in the first row of Figure 1. We can see that there is no clear regular orderliness for $t = 1$, which is the case that is most frequently adopted in the analysis of networks. However, the transparent in-degree layers gradually emerge when t goes large, and the difference of altitudes of layers become significant with t approaching n . We colorize the HI figure according to clusters and noise, as shown in Figures 3 (d) and (g). It is clear that the higher the density of cluster, the nearer the associated in-degree layer approaches the y axis. Moreover, these in-degree layers differentiate according to the distribution of cluster densities. We call the phenomenon of aggregation of in-degrees as the in-degree homophily. For much clearer illustration, we present the complete process of deformation of in-degree homophily in Video 1 of Supplementary Material¹.

Interestingly, the homophily in social science was vividly described as “birds of a feather flock together” (McPherson et al., 2001). For the geometric network here, we can clearly observe that the overall shape of the in-degree homophily in the HI figure is really like the wing of a bird. Refer also to Figure 3 (e) for a better example.

3. Homophilic Clustering

We can develop useful methods for clustering with the interesting characteristic of the in-degree homophily, including extraction of cluster cores, detection of cluster-to-noise boundary, and algorithms of clustering with low price.

3.1. Cluster-Core Extraction

Clustering will be easy if we can accurately locate the core of each cluster. With the HI figure, this problem becomes simple to handle. We single out the data points whose t -order in-degrees are larger than the corresponding out-degrees. If seen from the HI figure, these are data points whose in-degrees lie above the out-degree curve. As Fig-

¹ All the supplementary materials of this paper are available at <http://sites.google.com/site/zhaodeli/>

Algorithm 1 t -Order Dual Degrees**Input:** The graph matrix W and the integer t .

- 1: Initialization. $d^{in} \leftarrow \mathbf{1}$ and $d^{out} \leftarrow \mathbf{1}$.
- 2: **for** $tt = 1$ **to** t
- 3: $d^{out} \leftarrow W d^{out}$ and $d^{in} \leftarrow W^\top d^{in}$.
- 4: $a = \frac{1}{2}(d^{in} + d^{out})^\top \mathbf{1}$.
- 5: $d^{in} \leftarrow d^{in}/a$ and $d^{out} \leftarrow d^{out}/a$.
- 6: **end**

Output: d^{in} and d^{out} .**Algorithm 2** HI Figure**Input:** The t -order d^{in} and d^{out} .The index vector $s = [1, \dots, n]$.

- 1: $\vec{d}^{out} \leftarrow \text{sort}(d^{out})$ in descending order:
Record the associated index order \vec{s} .
- 2: Order d^{in} by \vec{s} , $\vec{d}^{in} \leftarrow d^{in}(\vec{s})$.

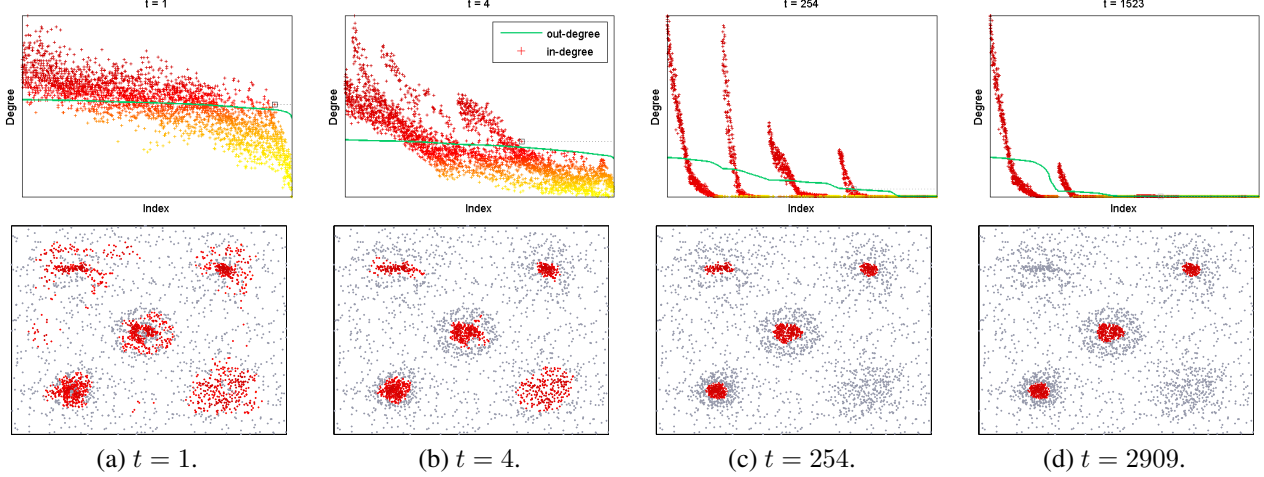
Output: Figure: $\text{plot}(s, \vec{d}^{out})$; $\text{plot}(s, \vec{d}^{in})$.

Figure 1. Toy example of the homophilic in-degree figure (HI figure). The 2D data contain five clusters of different densities in heavy noise, and the five clusters are nested with small clusters. Figures in the first row are the HI figures, while those in the second row are the corresponding cluster cores. The values of t are selected from the jump transitions in Figure 2 (a).

ures 1 (a)-(d) show, the separated points exactly consist of the cores of clusters for a moderate t . It attains messy results in the case of $t = 1$. The extracted cores become well-shaped with the growth of in-degree homophily over t . This critical clue leads us to define the homophilic coefficient for each node by

$$\bar{h}_i = \frac{d_i^{in}}{d_i^{out}}. \quad (3)$$

The homophilic coefficient \bar{h}_i of node i measures the degree that this node aggregates to be a member of a cluster. The larger the homophilic coefficient, the more important the node is from the clustering perspective. Therefore, we take the cluster cores out from noise by using $\bar{h}_i \geq 1$ for a proper t .

The homophilic coefficient of order 1 was previously proposed and applied for the detection of communities in the complex networks of the internet, genes, etc. (Maslov & Sneppen, 2002; Radicchi et al., 2004). For our geometric network, however, the 1-order homophilic coefficient fails to measure the popularity of clusters. However, we can also observe that the weak layers successively decay below the

out-degree curve with the increase of t , meaning that the cores of density-weak clusters are percolated out. Therefore, we need to formulate rules to attain an applicable t .

3.2. Cluster-to-Noise Boundary

The homophilic in-degree layer containing noise is easily identified due to the fact that the connection of network formed by noise is relatively weak. This means that the noise layer always lies at the tail of the HI figure. Another portion of noise consists of the bottom base of the HI figure, which are yielded by connections between clusters and noise. Therefore, the noise layer will be the first one to decay below the out-degree curve. When the noise layer disappears, the second-weakest layer will slither towards the tail of the HI figure over t until the growth of homophily converges. This dynamics of in-degree layers can be clearly observed from Video 1. In this way, we obtain a cue that quantitatively describes the deformation of in-degree layers. To do this, we define the residual distance of the HI figure: it is the distance between the right y axis to the point above the out-degree curve that is vertically nearest to it. We mark the points with gray squares and

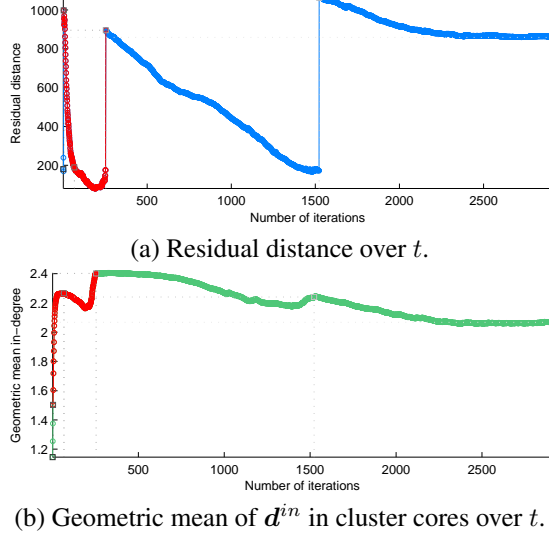


Figure 2. Residual distance and geometric mean of truncated in-degrees. The figures are best viewed in color and large size.

draw the distance paths with dotted lines in Figure 2. The trajectory of the residual distance over t is depicted in Figure 2, where we see that there is a jump transition when a weak layer decays. Therefore, we extract the largest cores of all clusters at the time at which the last point in the noise layer was just percolated by the out-degree curve. This pivotal time for the toy example is $t = 4$ at the first peak of jump transitions. The corresponding HI figure is Figure 1 (b). We denote the set of largest cores by \mathcal{C}_{\max} . To discriminate noise from clusters, we need to guarantee that the local density of any member in \mathcal{C}_{\max} is larger than all the members in $\mathcal{C}_{\text{noise}}$, where $\mathcal{C}_{\text{noise}}$ denotes the set of noise. Thus, we conclude the criterion of identifying the boundary between clusters and noise. Formally, we define the local density η_i of \mathbf{x}_i by the average of similarities in \mathcal{N}_i , i.e., $\eta_i = \frac{1}{k} \sum_{p=1}^k \text{sim}(\mathbf{x}_i, \mathbf{x}_{i_p})$. Investigating the distance or similarity of \mathbf{x}_i to its k -th NN is a general way of estimating local density of \mathbf{x}_i (Byers & Raftery, 1998). Here we use the average to enhance the robustness of estimator. Let

$$\eta_{\mathcal{C}_{\max}}^{\min} = \arg \min_{\mathbf{x}_i \in \mathcal{C}_{\max}} \eta_i. \quad (4)$$

The set of points in clusters can then be detected by

$$\mathcal{C}_{\text{cluster}} = \{\mathbf{x}_i | \eta_i \geq \eta_{\mathcal{C}_{\max}}^{\min}, \mathbf{x}_i \in \mathcal{X}\}. \quad (5)$$

The separated clusters from noise shown in Figure 3 (c) demonstrate that $\eta_{\mathcal{C}_{\max}}^{\min}$ is an effective estimator of cluster-to-noise boundary.

3.3. Determination of Powers

To extract better cluster cores, we must further determine another t . The interval between the first and second jump

transitions is the feasible set in which the selected t will produce the complete cores, because each cluster has points above the out-degree curve in this interval. We mark the feasible interval of determining t by red circles in Figure 2 (a). An optimal t for singling out cores should yield the optimal homophilic layers. Thus, a natural criterion is that the truncated in-degree layers by $\bar{h}_i \geq 1$ is maximally uniform, in the sense that the difference between the truncated in-degree layers of dense clusters and that of sparse clusters is minimized. By this criterion, we can select the balanced cores for all clusters, which is more favorable for clustering. A simple measurement for this optimality is the geometric mean of truncated in-degrees by $\bar{h}_i \geq 1$, showing that

$$g_t = \left(\prod_{\bar{h}_i \geq 1} d_i^{\text{in}} \right)^{\frac{1}{|\mathcal{C}_{\text{core}}^t|}}, \quad (6)$$

where $\mathcal{C}_{\text{core}}^t$ is the set of data points satisfying $\bar{h}_i \geq 1$. Figure 2 (b) illustrates the curve of g_t . The growth of strong layers and the reduction of weak layers shape the g_t curve with local maxima and minima. An optimal t we expect is at the local maxima in the feasible interval. The selected cores on toy data are shown in Figure 3 (a) and the associated HI figure in Figure 3 (e).

3.4. Homophilic Clustering

3.4.1. PAIR MERGING

With the extracted clusters and cluster cores, one can develop diverse approaches for clustering. Here we present a simple method based on the homophily-guided mergence of nodal links. Denote the set of extracted cores by $\mathcal{C}_{\text{core}}^t$, where t is optimally determined. For each $\mathbf{x}_i \in \mathcal{C}_{\text{core}}^t$, we take its k_c nearest neighbors. Here k_c is a small constant, generally, $k_c \in [1, 5]$. For the ideal case, we can directly merge these selected nearest neighbors in $\mathcal{C}_{\text{core}}^t$ to get clusters of cores if they are connected. For complex data, however, there may still be links between cores, which may deteriorate clustering results. To maintain the robustness to noisy links, we define a homophily-weighted similarity $\text{hsim}(\mathbf{x}_i, \mathbf{x}_j)$ between selected k_c NNs for cluster merging, giving that

$$\text{hsim}(\mathbf{x}_i, \mathbf{x}_{i_p}) = \bar{h}_i \bar{h}_{i_p} \text{sim}(\mathbf{x}_i, \mathbf{x}_{i_p}), \quad (7)$$

if $\mathbf{x}_i, \mathbf{x}_{i_p} \in \mathcal{C}_{\text{core}}^t$ and $\mathbf{x}_{i_p} \in \mathcal{N}_i^{k_c}, p = 1, \dots, k_c$. The $\text{hsim}(\mathbf{x}_i, \mathbf{x}_j)$ is the pairwise similarity weighted by the homophilic coefficients of associated NN pairs. This constraint ensures that the priority of messages passing is already along paths of the high homophily, thereby making the merging procedure robust to noisy links. With homophilic similarity, we can merge data pairs one by one from the largest $\text{hsim}(\mathbf{x}_i, \mathbf{x}_j)$ to the smallest one if they

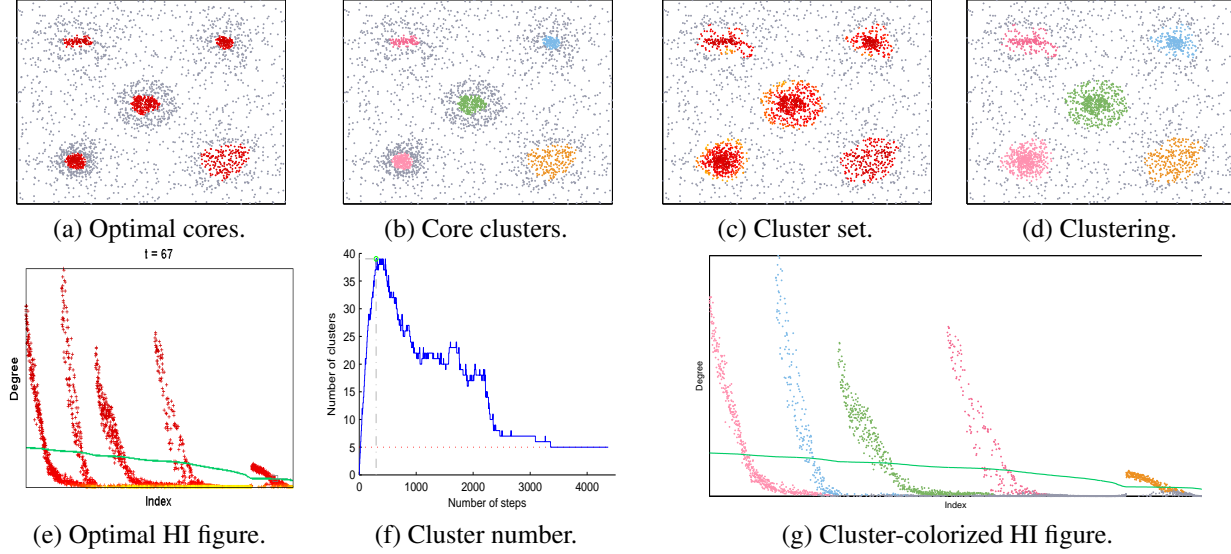


Figure 3. Core extraction and clustering. (a) and (b) are the extracted cores and the corresponding HI figure at $t = 67$, respectively. This value is determined by the optimal condition in Section 3.3. (b) is the clustering result of cores by homophily-guided merge. (c) is the complete set of clusters determined by formula 5. (d) Clustering by attachment in formula 9. (f) is the curve of the number of clusters when merging cluster cores. We set $k_c = 5$ for toy data. (g) is the optimal HI figure colorized by colors of clusters. These figures are best viewed in color and large size.

share mutual links, until the procedure converges or a given number of clusters is identified.

Figure 3 (b) shows the detected clusters of cores and Figure 3 (f) is the curve of the number of clusters during pair merging. The merging procedure converges when the cluster number c coincides with the real one of clusters.

3.4.2. AGGREGATION TO CORES

Let the resulting clusters of cores be denoted by $\mathcal{C}_{\text{core}}^t = \{\mathcal{C}_1, \dots, \mathcal{C}_c\}$. We need to assign the remaining data points in $\mathcal{C}_{\text{cluster}} \setminus \mathcal{C}_{\text{core}}^t$ to $\mathcal{C}_{\text{core}}^t$. We propose applying the leave-one-out strategy for assignment. The structural affinity of a point x_i to a cluster \mathcal{C}_j can be quantized by the variational value of its rank if we leave \mathcal{C}_j out from $\mathcal{C}_{\text{cluster}}$. For our framework, the ranks of x_i are in-degrees and out-degrees of order t . Therefore, we can investigate the ratio of $d_{i|\{\mathcal{C}_{\text{cluster}} \setminus \mathcal{C}_j\}}^{\text{in}}$ to $d_{i|\{\mathcal{C}_{\text{cluster}}\}}^{\text{in}}$, where the general expression $d_{i|\mathcal{C}}^{\text{in}}$ means the degree rank of x_i on \mathcal{C} . We compute the same ratio for $d_{i|\mathcal{C}}^{\text{out}}$. Putting these two dual ranks together, we derive the similarity measure of point-to-cluster affinity by product, writing it as

$$\rho_{x_i \rightarrow \mathcal{C}_j} = 1 - \frac{d_{i|\{\mathcal{C}_{\text{cluster}} \setminus \mathcal{C}_j\}}^{\text{in}}}{d_{i|\{\mathcal{C}_{\text{cluster}}\}}^{\text{in}}} \frac{d_{i|\{\mathcal{C}_{\text{cluster}} \setminus \mathcal{C}_j\}}^{\text{out}}}{d_{i|\{\mathcal{C}_{\text{cluster}}\}}^{\text{out}}} = 1 - \frac{\gamma_{i|\{\mathcal{C}_{\text{cluster}} \setminus \mathcal{C}_j\}}}{\gamma_{i|\{\mathcal{C}_{\text{cluster}}\}}}, \quad (8)$$

where $\gamma_{i|\mathcal{C}} = d_{i|\mathcal{C}}^{\text{in}} d_{i|\mathcal{C}}^{\text{out}}$ is the product rank of x_i . The larger the value of $\rho_{x_i \rightarrow \mathcal{C}_j}$ is, the more preference x_i has of being attached to \mathcal{C}_j . Therefore, the cluster label of x_i can be

inferred by

$$\arg \max_j \rho_{x_i \rightarrow \mathcal{C}_j}, j = 1, \dots, c. \quad (9)$$

Another benefit of applying the ratio of dual degrees to define $\rho_{x_i \rightarrow \mathcal{C}_j}$ is that the ratio can diminish the negative effect of inferring similarity caused by large degrees. The result of attaching toy data to cluster cores is shown in Figure 3 (d). To see the correspondence between clusters and homophilic layers, we colorize the HI figure according to the labels of the detected clusters and depict it in Figure 3 (g).

3.4.3. COMPLEXITY

It is straightforward to know that the space complexity of homophilic clustering is $\mathcal{O}(nk_c)$. In practice, k_c is a small integer in $\{1, 2, 3, 4, 5\}$. Thus, the complexity reduces to a linear one of $\mathcal{O}(n)$. The time complexity of homophilic clustering depends on the cluster structures of data points. Assume that the maximum number of clusters is c_{max} during pair merging and the corresponding number of iterations is m_{max} . c_{max} is actually determined by the connectivity of digraph \mathcal{G} and k_c . The time complexity is $\mathcal{O}(c_{\text{max}}m_{\text{max}} + (c_{\text{max}} - c)(n_c k_c - m_{\text{max}}))$, where $n_c = |\mathcal{C}_{\text{core}}^t|$ is the number of data points in the extracted core clusters. If the given number c of cluster is large, c approaches c_{max} , the complexity will be approximately $\mathcal{O}(cm_{\text{max}})$. If c is small, the worst case will be $\mathcal{O}(n_c c_{\text{max}})$. Usually, n_c is a small fraction of n . c_{max} will be reached for a moderate number m_{max} of iterations. The curve of c_{max}

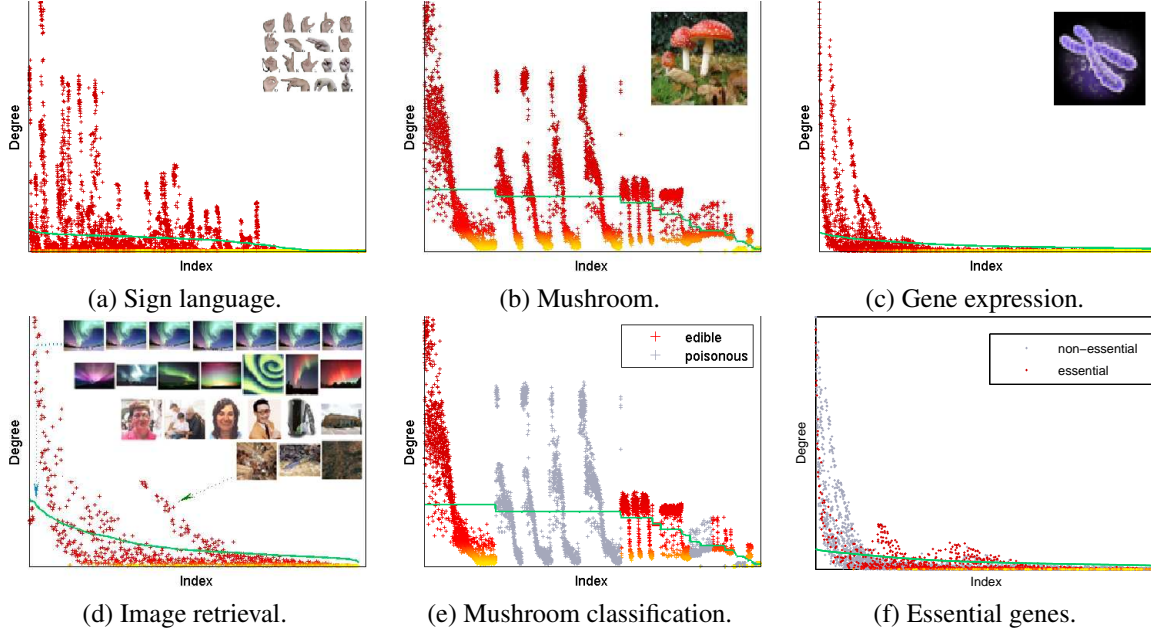


Figure 4. The HI figures on real-world scientific data. The optimal t for each data set is determined by the approach presented in Section 3.2. For the figure of gene data in (c), the optimal t is 45. For the figure of showing essential genes in (f), we carefully select the value of t to be 13 to highlight the homophilic layer of essential genes.

over m_{\max} on toy example is shown in Figure 3 (f) in the case of $k_c = 5$.

4. Experiment

We present more examples of the HI figures and compare our homophilic clustering algorithm with state-of-art algorithms on real-world data. The specific information of each data set is described in the supplementary material paper, which also contains the corresponding graph construction.

4.1. HI Figure

The homophilic effect of high-order in-degrees is also observed in real-world data from various scientific domains, as Figure 4 shows. Figure 4 (a) shows that the cluster densities of the hand-sign language data are very complicated, forming many homophilic layers. For Web images retrieved by the search engine, the clusters with clear semantics are detected in the HI layers, as Figure 4 (d) depicts. The semantically meaning images are contained in the strong HI layer, and the noisy images fall into the weakest HI layer and the base of the HI figure. An interesting observation is that many Web images of the same contextual content in the equal or different sizes are solely segregated to be a small agglomerate layer, which is presented by the vertical dotted arrow. This suggests the automatic filtering of redundant information for content-based image retrieval, which plays a central role in the next generation

of search engines. For the machine-intelligent discrimination of edible mushrooms from poisonous ones, the edible patterns exhibit transparently-layered regularity, providing considerable ease of classification, as shown by Figures 4 (b) and (e).

Of special scientific interest is the intriguing phenomenon observed from the kinetics of the HI layers of gene expressions in the budding yeast, *Saccharomyces cerevisiae* (Figure 4 (c)). In network biology, there has been lively debate in recent years concerning the spatial distribution of essential genes in functional modules of networks (Barabási et al., 2011). In light of our findings, a more elaborate structural organization of genes can be revealed from the HI figure. We have carefully checked the growth of the HI figure, and found a meaningful HI layer about essential genes around $t = 13$. Figure 4 (f) illustrates that only a small fraction of essential genes lie in the strong hub (core cluster) with the majority being peripheral. Interestingly, there are the two weak HI layers in which essential genes massively dominate. These two layers are so weak that they rapidly decay with the growth of the HI layers. This observation contributes to evidence that genes possess functional modules that are substantially composed of essential genes, and the sub-networks associated with these modules are very vulnerable, which is evident because the essential layers transiently exist. In addition, a considerable number of essential genes live in the base of the HI figure, implying that they are dispersively distributed outside functional

Table 1. Performance of clustering (%). The accuracy of clustering is measured by Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002). The abbreviations of the involved algorithms are listed as follows. Linkage, the average linkage algorithm. Zell, Zeta merging based on local links (Zhao & Tang, 2008). Ncuts, Normalized Cuts (Shi & Malik, 2000). SCK, spectral clustering with k-means (Ng et al., 2001). MCL, Markov Clustering (Dongen, 2000). AS, Authority shifting (Cho & Lee, 2010). HC, Homophilic Clustering (this paper). The ‘attribute’ refers to the most important feature of the category that the associated algorithm falls into. The denotation ‘-’ represents that the corresponding algorithms are computationally in-feasible for the data set.

Algorithmic Attribute	Partitional	Agglomerative		Spectral		Matrix power		
Algorithm	k-means	Average Linkage	Zell	Ncuts	SCK	MCL	AS	HC
FRGC	90.4	95	98.1	92.4	90.7	88.2	88.9	97.3
COIL	82.4	89.5	91	81.9	80.5	82.3	82.9	97.2
MNIST	54.6	-	-	63.9	66.9	-	-	81.4

communities. It is worth noting that these details are apparent only in the moderate evolution of dual degrees over t . Video 2 shows the complete dynamic process of the growth of the HI figure.

These examples verify that the HI figure can capture the intrinsic structures of data and is a powerful tool for data visualization and analysis.

4.2. Clustering

We perform the experiments of pattern clustering on the three widely applied benchmark databases in face recognition, object classification, and handwritten digit recognition. The face data are from the FRGC (Face Recognition Grand Challenge) database², which contains 466 persons (clusters) of 16,028 facial images. The number of members in each cluster varies from 2 to 80. The data set of object classification is the processed COIL database³, which contains 7,200 images of 100 objects. Each object cluster has 72 imagery members. The handwritten digits are from the well-known MNIST database⁴. The MNIST data set includes 70,000 handwritten digits of 10 classes. The algorithms we select to compare are presentative for clustering and most relevant to ours. For graph-based algorithms, we adopt the same directed graph for all algorithms, which can guarantee the fair comparison. We list the compared algorithms and the accuracy of each algorithm in Table 1.

Table 1 shows that on the relatively simple data like FRGC, the graph-theoretic methods based on hierarchical agglomerative clustering yield the best results and our HC performs comparably well. With the complexity of data increasing, the superiority of HC emerges. On the COIL data, HC is considerably better than the remaining algorithms. On MNIST, our algorithm significantly outperforms all the compared algorithms. The result of clustering the MNIST

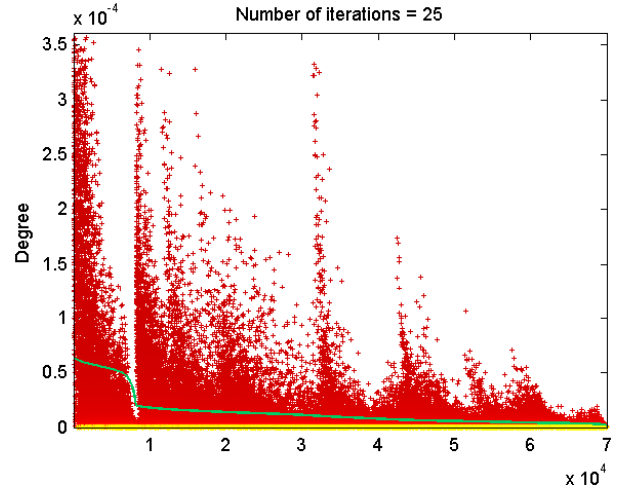


Figure 5. The HI figure on MNIST data ($t = 25$). For good visualization, we truncate the top part of the first layer to highlight the shapes of the remaining layers.

data proves the robustness of our algorithm to noisy data. Those algorithms of space complexity $\mathcal{O}(n^2)$ are computationally prohibitive for the 70,000-scale MNIST data. The HI figure of MNIST data is shown in Figure 5 for interested readers’ reference. Note that the single linkage algorithm can be scaled to cluster the 70,000 MNIST data. However, its accuracy on MNIST is too low and much lower than the average linkage algorithm on the other three datasets. So we show the results of the average linkage.

5. Conclusion

We have reported an interesting property of geometric digraph drawn from neighborhood asymmetries of data. The similarity propagation of local asymmetries leads to the homophilic distribution of in-degrees. Based on this finding, we have proposed an approach called the homophilic in-degree figure to data visualization and developed an algorithm to detect clusters from heavy noise. Extensive ex-

²<http://www.frvt.org/FRGC/>

³<http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php>

⁴<http://yann.lecun.com/exdb/mnist/>

periments on toy data and real scientific data validated the effectiveness of our algorithms. In addition to the applications in pattern clustering, our algorithms can also be applicable for vector quantization, Nyström matrix approximation, topic models, and image segmentation, in which cases clusters play an important role.

Acknowledgement

We are aware that a paper published on Science very recently (Rodriguez & Laio, 2014) handles the similar problem of clustering with the one presented in this paper.

References

- Barabási, A.L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- Barabási, A.L., Gulbahce, N., and Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68, 2011.
- Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Journal of Neural Computation*, 15:1373–1396, 2003.
- Byers, S. and Raftery, A.E. Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of The American Statistical Association*, 93(442):577–584, 1998.
- Chiang, K.Y., Natarajan, N., Tewari, A., and Dhillon, I.S. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM international conference on Information and Knowledge Management (CIKM 2011)*, pp. 1157–1162, 2011.
- Cho, M. and Lee, K.M. Authority-shift clustering: hierarchical clustering by authority seeking on graphs. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010.
- Dongen, S. Van. *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, 2000.
- Frey, B.J. and Dueck, D. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.
- Kandola, J., Shawe-taylor, J., and Cristianini, N. Learning semantic similarity. In *Advances in Neural Information Processing Systems (NIPS 2003)*, Cambridge, MA, 2003. MIT Press.
- Katz, L. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.
- Kossinets, G. and Watts, D.J. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115(2):405–450, 2009.
- Lin, F. and Cohen, W.W. Power iteration clustering. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 655–662. ACM press, 2010.
- Maslov, S. and Sneppen, K. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.
- McPherson, M., Smith-Lovin, L., and Cook, J.M. Birds of a feather: homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.
- Meilă, M. and Pentney, W. Clustering by weighted cuts in directed graphs. In *In Proceedings of the 2007 SIAM International Conference on Data Mining (SDM 2007)*, 2007.
- Newman, M.E.J. and Girvan, M. Finding and evaluating community structure in networks. *Physics Review*, 69:167–256, 2003.
- Ng, A.Y., Jordan, M.I., and Weiss, Y. On spectral clustering: analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS 2001)*, Cambridge, MA, 2001. MIT Press.
- Papadopoulos, F. and Kitsak, M. Popularity versus similarity in growing networks. *Nature*, 489(7417):537–540, 2012.
- Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. Defining and identifying communities in networks. In *Proceedings of the National Academy of Sciences (PNAS 2004)*, pp. 2658–2663, 2004.
- Rodriguez, A. and Laio, A. Clustering by fast search and find of density peaks. *Science*, 344:1492–1496, 2014.
- Shi, J.B. and Malik, J. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 22(8):888–905, 2000.
- Strehl, A. and Ghosh, J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.
- Tenenbaum, J.B., Silva, V., and Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.
- Zhang, W., Wang, X.G., Zhao, D.L., and Tang, X.O. Graph degree linkage: agglomerative clustering on a directed graph. In *Proceedings of European Conference on Computer Vision (ECCV 2012)*, 2012.

Zhao, D.L. and Tang, X.O. Cyclizing clusters via zeta function of a graph. In *Advances in Neural Information Processing Systems (NIPS 2008)*, pp. 1953–1960, Cambridge, MA, 2008. MIT Press.

Zhou, D., Huang, J., and Schölkopf, B. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the 22th International Conference on Machine Learning (ICML 2010)*, pp. 1041–1048. ACM press, 2005.